

# Application of Principal Component-Minimum Variance Technique in Gene Prediction

M. Roy<sup>1</sup>, S. Barman (Mandal)<sup>2</sup>

The Calcutta Technical School, Govt. of W.B.<sup>1</sup>

Institute of Radio Physics & Electronics, University of Calcutta.<sup>2</sup>

110, S.N. Banerjee Road, Kolkata-700013, India.<sup>1</sup>

92, A.P.C. Road, Kolkata-700 009, India.<sup>2</sup>

<sup>1</sup>dipamani\_ccp@rediffmail.com, <sup>2</sup>barmanmandal@gmail.com

## Abstract

The problem under consideration concerns information extraction from eukaryotic DNA sequences regarding existence of protein coding regions. Spectral Analysis using classical Fourier Transform techniques such as Discrete Fourier Transform (DFT) has long been used for this purpose with the help of period-3 peaks. Since this method has low Signal to Noise Ratio (SNR), the spectral peaks are difficult to distinguish in the background of noise. Researchers have designed various types of filters to suppress this noise so that the period-3 peaks are revealed prominently. In this article, a novel approach was applied by combining Principal Component Analysis with Minimum Variance Estimator for effective gene prediction. Here PSD of DNA sequence has been estimated using Minimum Variance method in which noise reduction has been accomplished by Principal Component Analysis of correlation matrix. In the process, the dimension of the data-set was reduced by projecting the raw data onto a few prominent eigenvectors with large eigenvalues. The resulting reduced-rank approximation to correlation matrix was then used for spectrum estimation. The results were compared with those of Blackman-Tukey Power Spectrum Estimator which is a modified form of Periodogram method. Eukaryotic genes from various organisms taken from NCBI Genbank have been used as test samples. A single sequence mapping method comprising real and imaginary values towards nucleotide bases was employed. The superiority of PCA based Minimum Variance method over Blackman-Tukey method was established with the help of spectral plots in perspective of both resolution and quality factor.

## Keywords

*Non-parametric; Periodogram; Power Spectral Density; Codon; Principal Component Analysis; Eigen-vector; Minimum Variance*

## Introduction

At the basic level genomic sequence information is discrete in nature because there are only a finite

number of nucleotides in the form of alphabets. In order to capture its periodic characteristics, one can interpret the DNA sequence as a discrete time sequence that can be analyzed using digital signal processing tools. Locating protein coding segments (exons) in DNA has been an important application in genomic area. FFT based power spectrum estimation methods known as classical methods have been widely used for this purpose. In this paper, the Minimum Variance(MV) technique has been employed for Power Spectral Density(PSD) estimation of a DNA sequence incorporating Principal Component Analysis(PCA) of correlation matrix for noise removal. The power spectral peaks are used as identification features for coding regions. PCA is one of the most valuable tools of linear algebra which is used extensively in all forms of analysis because it is a simple non-parametric method of extracting relevant information from confusing data-sets, and provides a roadmap to reduce complex dataset to a lower dimension so that hidden periodicities can be revealed.

It is well known that genetic information is stored in the particular order of four kinds of nucleotide bases, Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) which comprises the DNA bio-molecule along with Sugar-Phosphate backbone. DNA sequence is classified into exons (coding region) and introns (non-coding region). Exons of a DNA sequence are the most information bearing part because only the exons take part in protein coding while the introns are spliced off during protein synthesis.

Gene prediction refers to detecting locations of the protein coding regions of genes in a long DNA sequence. There has been great deal of work done in applying Digital Signal Processing and Statistics

methods to DNA in the recent past, some of which are mentioned here in brief. Its application in finding periodicities in DNA sequences has been explored by various researchers (Anastassiou,2000 and 2001; Zhao, 2006). It has been established that base sequences in the exon regions of DNA molecules exhibit a period-3 property because of the codon structure involved in the translation of nucleotide bases into amino acids (Fickett and Tung,1982; Tiwari, Ramachandran, Bhattacharya, Bhattacharya and Ramaswamy,1997; Yin, Stephen and Yau,2007). Investigation into long range correlation has also been the focus of attention for many researchers (Li and Kaneko,1992; Voss,1992; Herzel and Grobe,1995). A coding measure scheme employing electron-ion-interaction pseudo-potential (EIIP) was presented as a revision for binary indicator sequences (Nair and Sreenadhan,2006; Mabrouk,2012). Implementation of digital filters to extract period-3 components and that effectively eliminate background 1/f noise present in DNA sequence has given good results (Vaidyanathan and Yoon,2004; Tuqan and Rushdi,2008). Positional Frequency Distribution of nucleotides presented by Roy, Biswas and Barman (Mandal) (2009) has given interesting results. Spectral analysis of coding and non-coding regions of a DNA sequence by parametric and non-parametric methods has been discussed by Roy and Barman (Mandal) (2011) with the goal of gene detection. In the present paper, the authors have identified coding regions of eukaryotic DNA from several organisms based on PSD plots using PCA based Minimum Variance estimator and compared the results with Blackman-Tukey estimator (Hayes,1996). DNA (de- oxyribo-nucleic acid) is a huge data-base available in Public Domain having hereditary traits hidden in it (<http://www.ncbi.nlm.nih.gov>).

The paper is organized into different sub-sections. In the first three sub-sections of section-2, mathematical background of non-parametric methods is stated and flow-chart of the proposed algorithm is given in the last sub-section. Results obtained from various techniques are analyzed in section-3. Finally, conclusion drawn is mentioned in section- 4.

## PSD Estimation of DNA Sequences

DSP can be used as an effective tool to analyze the vast genomic data available in the NCBI Genbank. DSP technique is applicable only to numerical data hence the four-letter alphabetic sequence has to be mapped into numerals before applying DSP tools. Hence different mapping methods have been adopted for

this purpose. Here the authors have applied a quaternary mapping rule assigning:

$$a=-1, c=-j, g=1 \text{ and } t=j.$$

Non-parametric technique of PSD estimation using Fourier Transform known as Periodogram can be classified as direct and indirect. The direct method takes Discrete Fourier Transform (DFT) of the signal and then averages the square of its magnitude. The indirect method is based on the idea of first estimating the autocorrelation of data sequence then taking its Fourier Transform. Due to poor quality of periodogram as an estimator of PSD, various improved versions such as Welch, Bartlett, Blackman-Tukey etc have been developed. In the process, signal is divided into overlapping segments, each data segment is windowed, periodogram is estimated and averaged. The other type of non-parametric method discussed in this paper is the Minimum Variance method proposed by Capon. This spectrum estimator can be interpreted as a bank of data-adaptive narrow-band FIR band-pass filters which are optimized to minimize their response to components outside their band of interest. The MV estimator has better resolution than Periodogram and BT estimator. Here the authors used Minimum Variance technique to estimate PSD in a novel way by combining it with Principal Component Analysis for effective prediction of protein coding regions in DNA.

## Blackman-Tukey Method of Power Spectrum Estimation

In Blackman-Tukey(BT) method of power spectrum estimation, the Discrete Time Fourier Transform (DTFT) of a windowed autocorrelation sequence  $r_x(k)$  of process  $x(n)$  with lags  $|k| \leq M$  is taken. Process  $x(n)$  signifies nucleotide test data of length  $N$  mapped into numerical form.

$$P_{BT}(e^{jw}) = \sum_{k=-M}^M r_x(k)w(k)e^{-jwk} \quad (1)$$

Here the estimated autocorrelation function is given by:

$$r_x(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} x(n+k)x^*(n) \quad (2)$$

Where  $k=0,1,2,\dots,(N-1)$ , with  $r_x(k)$  set equal to 0 for  $|k| \geq N$ . With values of  $r_x(k)$  for  $k < 0$  defined using conjugate symmetry given as:  $r_x(-k) = r_x^*(k)$ .

If  $w(k)$  is a Bartlett (triangular) window then BT estimate is written in terms of autocorrelation as:

$$P_{BT}(e^{jw}) = 1/M \sum_{k=-M}^M (M - |k|) r_x(k) e^{-jwk} \quad (3)$$

### Principal Component Spectrum Estimation:

This algorithm is based on Principal Component Analysis of correlation matrix. After eigen-decomposition of the correlation matrix, the eigen-values  $\lambda_i$  are arranged in descending order:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_M$$

The prime eigen-values of dimension  $p$  with steep slope correspond to signal subspace and the set of smallest eigen-values of dimension  $(M-p)$  having more or less flat nature correspond to noise subspace. Hence rank  $p$  constraint is imposed on the correlation matrix effectively filtering out the noise subspace thus enhancing the signal components. Let  $R_x$  be an  $M \times M$  autocorrelation matrix of the signal consisting of  $p$  complex exponentials in white noise. The eigen-decomposition of  $R_x$  is given by:

$$R_x = \underbrace{\sum_{i=1}^p \lambda_i \vec{v}_i \vec{v}_i^H}_{\text{signal}} + \underbrace{\sum_{i=p+1}^M \lambda_i \vec{v}_i \vec{v}_i^H}_{\text{noise}} \quad (4)$$

$H$  denotes the conjugate transpose,  $\vec{v}_i$  the  $i^{\text{th}}$  eigenvector of the autocorrelation matrix and  $\lambda_i$  the  $i^{\text{th}}$  eigen-value. On effectively filtering out noise portion, the spectral component due to signal alone is enhanced.

### Principal Component Minimum Variance Power Spectrum Estimation:

Given the autocorrelation sequence  $r_x(k)$  of a process  $x(n)$  for lags  $|k| \leq M$ . The  $M^{\text{th}}$  order estimate of Minimum Variance spectrum is given by:

$$P_{MV}(e^{jw}) = M/(\vec{e}^H R_x^{-1} \vec{e}) \quad (5)$$

Here  $\{\vec{e}\}$  is the vector of complex exponentials.

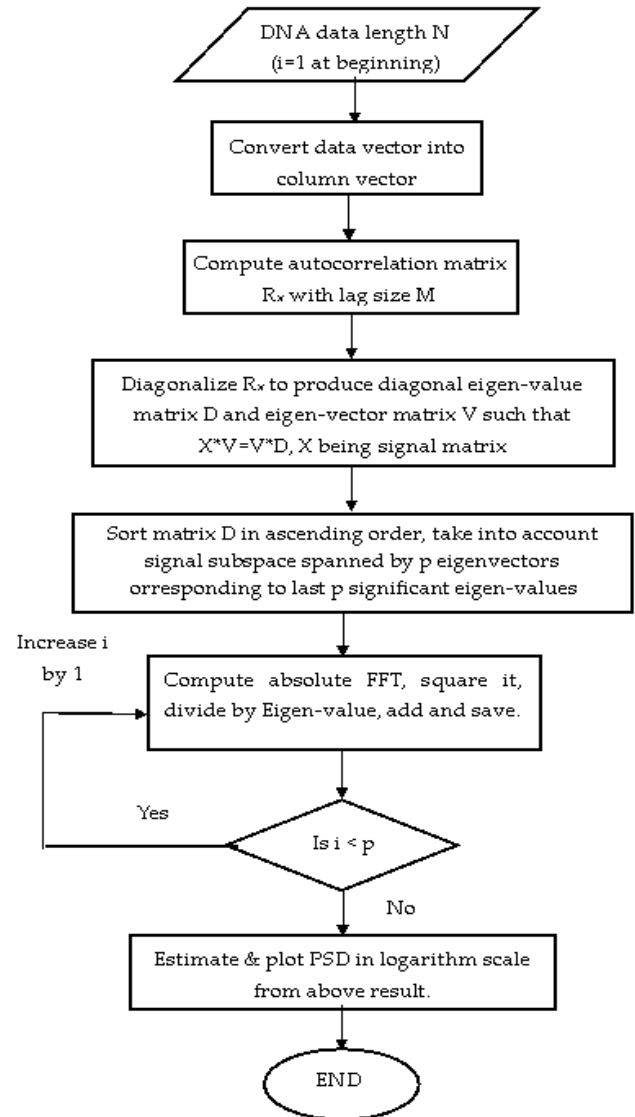
On eigen decomposition of the autocorrelation matrix, the inverse is given as:

$$R_x^{-1} = \sum_{i=1}^p \frac{1}{\lambda_i} \vec{v}_i \vec{v}_i^H + \sum_{i=p+1}^M \frac{1}{\lambda_i} \vec{v}_i \vec{v}_i^H \quad (6)$$

Where  $p$  is the number of complex exponentials retaining the first  $p$  Principal Components of  $R_x^{-1}$  giving the Principal Component MV estimate:

$$P_{PC-MV}(e^{jw}) = M / \sum_{i=1}^p \frac{1}{\lambda_i} |\vec{e}^H \vec{v}_i|^2 \quad (7)$$

### Flowchart of Algorithm for proposed Principal Component Minimum Variance PSD Estimator:



### Simulation Results and Discussion

In this paper, two classical techniques of Power Spectral Density Estimation have been explored in MATLAB version 7.1 environment to find coding segments in DNA sequences effectively. The authors have presented a comparative study of PSD plots of DNA sequences from a standard non-parametric approach known as Blackman-Tukey with that of proposed PCA based Minimum Variance approach. The nucleotide bases from various organisms have been used as raw data for analysis. TABLE-1 indicates the DNA length, location and length of exons and other details of the organisms tested. It is known that PSD describes how the average power of a signal  $x[n]$  is distributed with frequency, where  $x[n]$  is a sequence of random variables defined for every integer  $n$ . According to period-3 property of DNA, there must be

prominent visible peaks in the exon areas in PSD plots. Among all popular classical spectrum estimation techniques based on Fourier Transform Blackman-Tukey technique is one of the most improved version of periodogram giving best statistical properties. It provides averaged and smoothed periodogram with segmented, windowed and overlapped signal ( Fig. 1A - Fig. 7A ). In the present work, an overlap of 50% with Bartlett (triangular) window of suitable segment lengths each has been used (TABLE-2). The choice of window length  $M$  has been done subjectively based on a trade-off between spectral resolution and statistical variances. If  $M$  is too small important features may be smoothed out while if  $M$  is too large the behaviour becomes more like periodogram with erratic variation. Therefore, a compromise value is chosen between range  $1/25 < M/N < 1/3$ , where  $N$  is nucleotide sequence length. All the PSD BT plots (FIG.1A-FIG.7A) show period-3 peaks in accurate locations in presence of background noise often making it difficult to locate the actual period-3 peak corresponding to the exon segment.

Next a novel method was utilized known as Principal Component Minimum Variance(PC-MV) PSD estimator using concept of Principal Component Analysis of correlation matrix. In this process, the dimension of the dataset is reduced by projecting the raw data onto a few prominent eigenvectors with large eigen-values. A key issue in developing PCA model is to choose adequate number of principal components (PCs) to represent the system in an optimal way. Hence for reduction of dataset selection of proper model order is important in application of principal component analysis. A very simple method based on eigen-ratio has been adopted. As shown in FIG.8 Eigen-value Ratio  $\lambda_p / \lambda_{p+1}$  vs order  $p$  has been plotted for *Platisthys Flesus* gene. It is observed that there has an eigen-value gap of high magnitude between orders  $p=14$  and  $p=15$ . This fact suggests that satisfactory estimate of rank of  $R_x$  is 14 for *Platisthys Flesus* gene. Thus it may be assumed that eigen-values  $\lambda_{15}$  onwards are the noise eigen-values which can be ignored (Liavas and Regalia, 2001).

The proposed algorithm was tested on various genes to predict location of coding regions of varying lengths and simulation results were compared with those of standard BT method on the same DNA data. It is seen that the new basis filters out the noise and reveals the hidden periodicities prominently with higher peak strength (Fig.1B-Fig.7B). The plots reveal that PC-MV spectrum estimate provides better prominence and

higher resolution than BT method. But the price that has to be paid is in computation time as mentioned in TABLE-2. An important metric for comparison of spectrum is the resolution of the PSD estimator, corresponding to the ability of the estimator to provide fine details of the PSD of data. The resolution and variance of the MV estimators depend on choice of filter order. If filter order is large, bandwidth of the filter is small and there is better rejection out of bound power providing better resolution. But larger filter order requires more autocorrelation lags in the autocorrelation matrix. This increases the computation complexity of estimating the autocorrelation matrix  $R_x$  and its inverse  $R_x^{-1}$  enhancing the computation time as well as the variance. Hence filter order is kept much below the length of data striking a trade-off between resolution and variance. The PC-MV plots (FIG.1B-FIG.7B) reveal sharp period-3 peaks in exon regions which can be located without any ambiguity though there are a few exceptions. For example, one can notice a small spurious peak in FIG.1B and a split in peak-3 in FIG.4B. In spite of this sort of limitations the overall performance of PC-MV technique over BT technique is far superior.

Another performance characteristic of the Power Spectral Estimator is Quality Factor which is a measure of ratio of mean square to variance of estimated PSD. TABLE-2 indicates the statistical parameters and computation times of both BT and PC-MV methods for genes *Bovine Gastrin*, *Platisthys Flesus*, *Didelphis Marsupialis*, *Ovis Aries*, *Cavia Porcellus*, *Caleglobim* and *Pigapai*. The bar plot of Quality Factors in Fig.9 indicates that Quality Factor obtained by PC-MV estimator is much higher than standard BT estimator.

Tomar, Gandhi and Vijaykumar(2008) in their work used Minimum Variance spectral estimator for gene prediction in an effort to detect small exons, and found that MV estimator outperformed other classical DSP tools like Fourier Transform etc to some extent. In this work, it has been observed that the PC-MV method makes further improvement providing much sharper period-3 peaks than methods proposed earlier even for small closely spaced exons though the price one has to pay is in computation time.

## Conclusion

In this paper, the power spectrum of several exon segments of eukaryotic genes belonging to a number of organisms has been checked using a single sequence indicator consisting of real and imaginary numbers.

The classical spectral estimation approach is computationally efficient but suffers from limited resolution problem. Hence it is required to look for a method that has better resolution giving prominent period-3 peaks. A novel and modern spectral estimation method known as Principal Component Minimum Variance method has been applied to DNA sequence and the results were compared with those from Blackman-Turkey Power Spectrum Estimator (PSE). It was noted that the standard non-parametric PSE methods are methodologically straight forward,

computationally simple and easy to understand. But it has low Signal to Noise Ratio (SNR) and spectral features are difficult to distinguish as noise artifacts appear in spectral estimates. Hence identifying Protein coding regions becomes difficult. Whereas application of Principal Component Analysis to correlation matrix for dimensionality reduction to Minimum Variance algorithm provides a PSD estimate which reveals sharp period-3 peaks in the coding regions of DNA sequence providing un-ambiguous exon prediction.

TABLE 1: DETAILS OF SAMPLE DATA-BASE

Gene Id	Gen Bank Accession No.	DNA Length In bp	Length of exons in bp	Source
BOVGAS	M31657.1	1066	315 (540-750, 896-999)	Bovine Gastrin
AF135499	AF135499.1	1845	1127(1-123, 228-467, 857-1295, 1408-1589, 1702-1845)	Platisthys Flesus (European flounder)
DMPROTP1	L17007.1	624	177 (122-248, 376-425)	Didelphis Marsupialis (Southern opossum)
OAMTTI	X07975.1	2055	186 (995-1022, 1312-1377, 1697-1788)	Ovis aries (sheep)
GPINCP1AA	D14119.1	2760	396 (1669-1840, 2402-2511)	Cavia Porcellus (Domestic Guinea pig)
CALEGLOBIM	L25363.1	1698	444 (144-235, 364-586, 1399-1527)	Callithrix Jacchus (White tufted ear marmoset)
PIGAPAI	L00626.1	3333	798 (751-793, 975-1128, 1770-2370)	Sus Scorfa (pig)

TABLE 2: SUMMARY OF STATISTICAL PARAMETERS &amp; COMPUTATION TIME OF BT METHOD AND PC-MV METHOD FOR VARIOUS GENES

Organisms	Blackman-Tukey (db)				PCA Based Minimum Variance (db)			
	Q.F. mean) <sup>2</sup> /var	Computation Time in Sec.	Window Length M	Segment No. K	Q.F. mean) <sup>2</sup> /var	Computation Time in Sec.	Order p	Lag Window M
BOVGAS	2.02	0.02	210	5	16.17	0.95	4	256
PLATISTHYS FLESSUS	3.24	0.17	205	9	12.54	34.89	14	1024
DMPROTP1	2.34	0.11	104	6	6.72	0.025	6	64
OAMTTI	6.50	0.38	205	10	12.49	1.56	7	256
GPINCP1AA	3.16	0.37	138	20	29.04	43.55	2	1024
CALEGLOBIM	2.46	0.09	212	8	25.77	6.53	3	512
PIGAPAI	2.39	0.28	303	11	29.84	45.38	4	1024

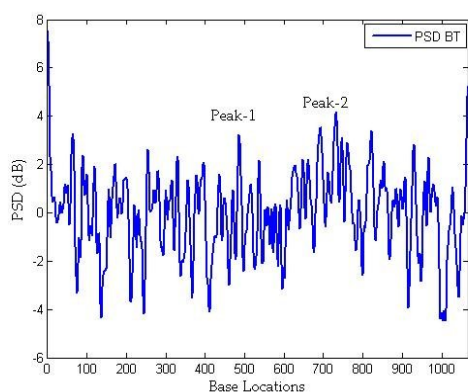


FIG.1A BOVINE GASTRIN GENE PSD BT

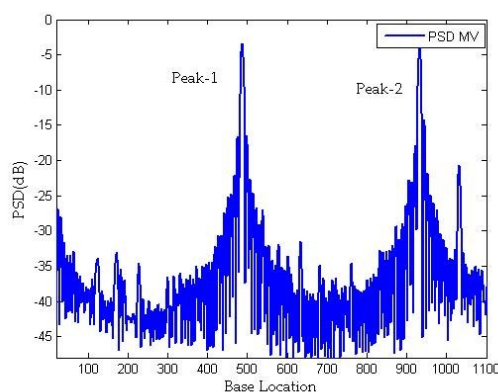


FIG.1B BOVINE GASTRIN GENE PSD PC-MV

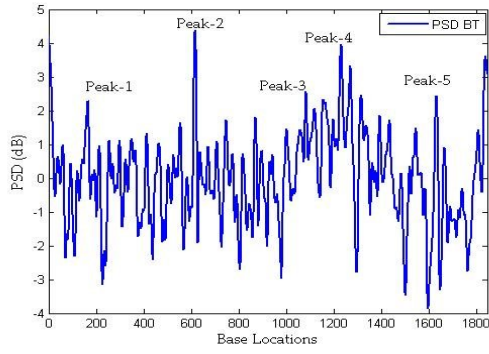


FIG.2A PLATICHTHYS FLESUS PSD BT

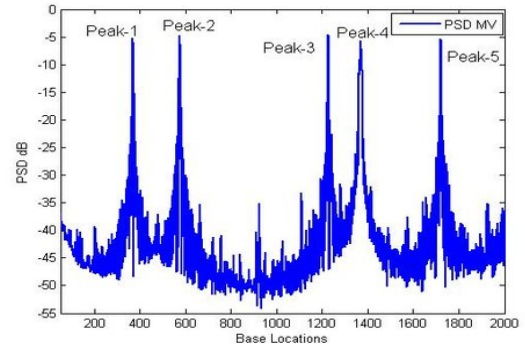


FIG.2B PLATICHTHYS FLESUS PSD PC-MV

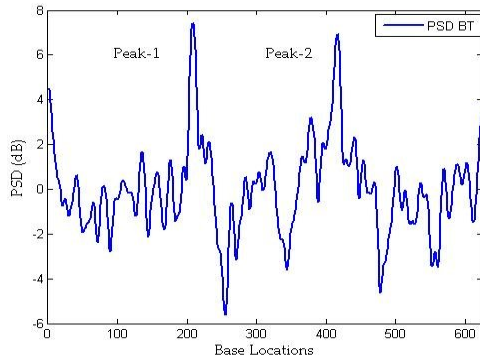


FIG.3A DIDELPHIS MARSUPIALIS PSD BT

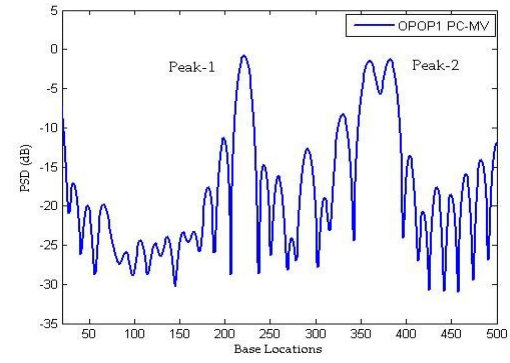


FIG.3B DIDELPHIS MARSUPIALIS PSD PC-MV

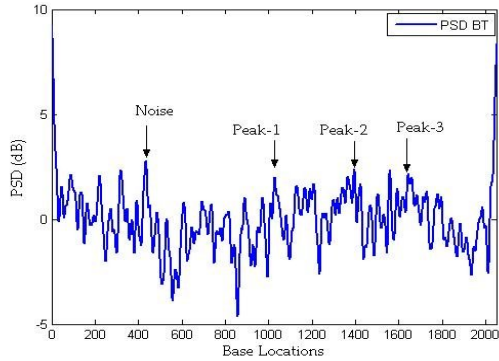


FIG.4A OVIS ARIES GENE PSD BT

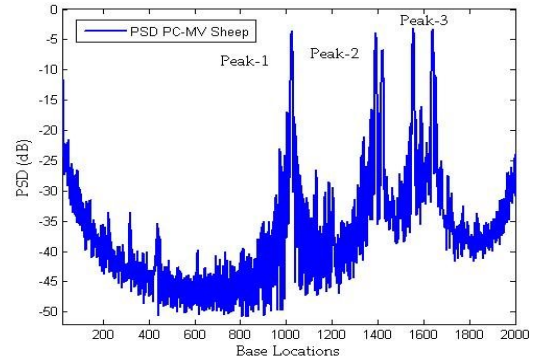


FIG.4B OVIS ARIES GENE PSD PCA MV

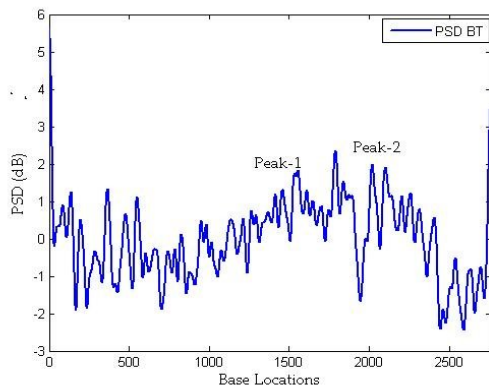


FIG.5A CAVIA PORCELLUS PSD BT

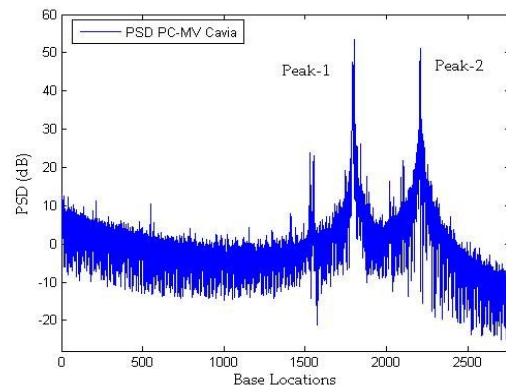


FIG.5B CAVIA PORCELLUS PSD PC-MV

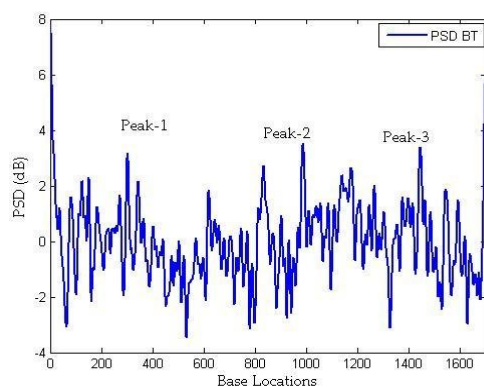


FIG.6A CALEGLOBIM GENE PSD BT

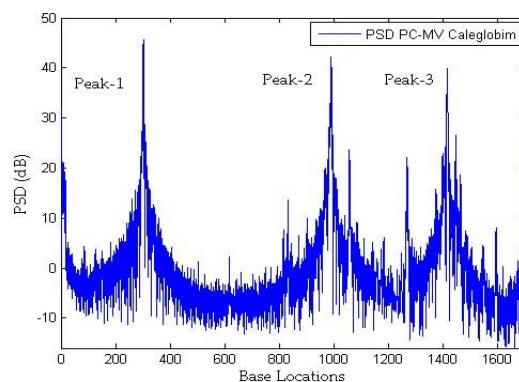


FIG.6B CALEGLOBIM GENE PSD PC-MV

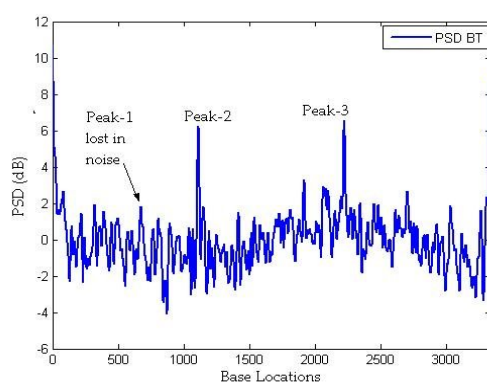


FIG.7A PIGAPAI GENE PSD BT

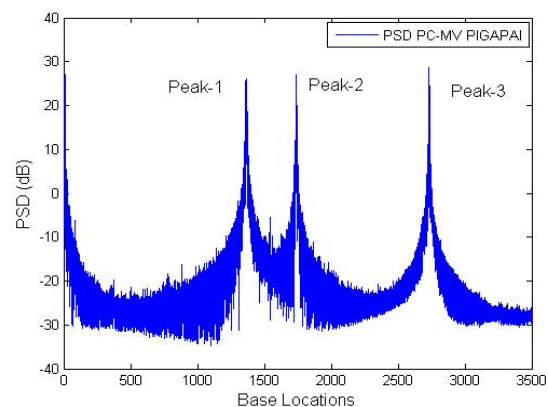


FIG.7B PIGAPAI GENE PSD PC-MV

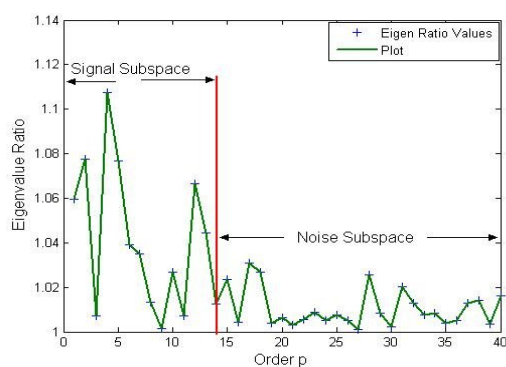


FIG.8 EIGEN RATIO PLOT FOR PLATICHTHYS FLEUSUS

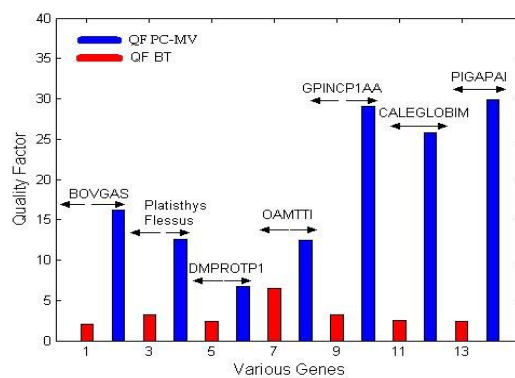


FIG.9 QF BAR PLOT FOR VARIOUS GENES

## REFERENCES

- Anastassiou, D., "DSP in genomics: processing and frequency domain analysis of character strings," IEEE, 0-7803-7041-2001.
- Anastassiou, D., "Frequency- domain Analysis of Biomolecular Sequences," Bioinformatics, 16 (2000): 1073-1081.
- Capon, J., "High- resolution frequency- wavenumber spectrum analysis," Proc., 57(1969): 1408-1418.

- Ficket, J. W. and Tung, C.S., "Recognition of protein coding regions in DNA sequences," Nucleic Acids Research, 10(17) (1982): 5303-5318.
- Hayes, M.H., Statistical Digital Signal Processing and Modeling, (John Wiley & Sons, Inc., New York, USA, 1996), 420-472.
- Herzel, H. and Grobe, B., "Measuring correlations in symbol sequences," Phys. A, 216 (1995): 518-542.
- Li, W. & Kaneko, K., "Long range correlation and partial 1/f

- spectrum in a non- coding DNA sequence," *Europhys. Lett.*, 17(7) (1992): 655-660.
- Liavas, A.P., Regalia, P.A., "On the Behavior of Information Theoretic Criteria for Model Order Selection," *IEEE Transaction on Signal Processing*, 49(8) (2001): 1689-1695.
- Mabrouk, M.S., "A Study of the Potential of EIIP Mapping Method in Exon Prediction Using the Frequency Domain Techniques," *American Journal of Biomedical Engineering*, 2(2) (2012): 17-22.
- Nair, A.S., Sreenadhan, S.P., "A coding measure scheme employing electron-ion-interaction pseudopotential (EIIP).," *Bioinformation*, I(6) (2006): 197-202.
- Roy, M and Barman, S. (Mandal), "Spectral analysis of coding and non-coding regions of a DNA sequence by parametric and non-parametric methods: A comparative approach", *Annals of Faculty Engineering Hunedoara, International Journal of Engineering, Romania, Fascicule-3* (2011): 57-62.
- Roy, M., Biswas, S. and Barman, S. (Mandal), "Identification and analysis of coding and non-coding regions of a DNA sequence by Positional Frequency Distribution of Nucleotides (PFDN) algorithm," Published in *Computers and Devices for Communication*, 2009, IEEE Xplore, ISBN: 978-1-4244-5073-2.
- Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R., "Prediction of probable genes by Fourier analysis of genomic sequences," *CABIOS*, 3(3) (1997): 263-270.
- Tomar, V., Gandhi, D. and Vijaykumar, C., "Digital Signal Processing for Gene Prediction ," *Proc. IEEE, TENCON(2008)*, Hyderabad, India, 18-21 Nov. 2008.
- Tuqan, J. and Rushdi, A., "A DSP based approach for finding the codon bias in DNA sequences," *IEEE Journal on Signal Processing*, 2 ( 3) ( 2008): 343-356.
- Vaidyanathan, P.P. and Yoon, B.J., "The role of signal-processing concepts in genomics and proteomics," *Journal of the Franklin Institute, Special issue on Genomics*, 341(2004): 111-135.
- Voss, R.F., "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Phy. Rev. Lett.*, 68(25) (1992): 3805-3808.
- Yin, C., Stephen, S. and Yau, T., "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence," *Journal of Theoretical Biology*, 247(2007): 687-694.
- Zhao, Lan, "Application of spectral analysis to DNA sequences," *CSD TR #06-003*, January 2006.
- <http://www.ncbi.nlm.nih.gov>



<sup>1</sup>**M.Roy** received her B.Sc. Engg. and M.E. degrees from BIT Mesra, Ranchi , India and Jadavpur University, Kolkata, India respectively in 1977 and 2003.

She had been teaching in Govt. Polytechnic till 2005. Presently, she is the Principal of The Calcutta Technical School, Govt. of West Bengal, Kolkata. Her research interests are Application of Digital Signal Processing in Genetics.



<sup>2</sup>**S.Barman**(Mandal) received B.Tech and M.Tech degrees in Radio Physics & Electronics from Calcutta University, India in 1994 and 1996 respectively and Ph.D degree from Jadavpur University, India in 2001.

She is an Assistant Professor of Institute of Radio Physics & Electronics, University of Calcutta, India. Her research interest includes Mechatronics, Digital Communication and Genomic Signal Processing & Modelling.